Modelo Lineal
PRACTICA 5 Parte 2

8. Considere el modelo $Y = X\beta + \epsilon$, donde $\epsilon \sim N_n(0, \sigma^2 I_n)$ y X es una matriz $n \times p$ de rango p. Sean $\lambda_1, \lambda_2, \ldots, \lambda_p$ los autovalores de la matriz $X'X$ y llamemos $\widehat{\beta}$ al estimador de mínimos cuadrados de β.

a) Muestre que $\sum_{j=1}^{p} Var(\widehat{\beta}_j) = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$. ¿Cuál es el efecto de pequeños autovalores sobre la varianza de los coeficientes?

b) Si $tr(X'X) = c$, donde c es una constante dada, pruebe que $\sum_{j=1}^{p} Var(\widehat{\beta}_j)$ se minimiza si $X'X = c/pI_p$.

c) Muestre que $E(\widehat{\beta}'\widehat{\beta}) = \beta'\beta + \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$. ¿Cuál es el efecto de pequeños autovalores sobre la tendencia de $\widehat{\beta}'\widehat{\beta}$ a sobreestimar $\beta'\beta$?

d) Suponga que podría incorporar m observaciones adicionales que provienen del modelo supuesto y que estas observaciones están contenidas en la matriz X^* de dimensión $m \times p$ con $m \geq p$ y además que $X^{*\prime}X^* = dI_p$. Pruebe que después de incorporar las m observaciones adicionales $\sum_{j=1}^{p} Var(\widehat{\beta}_j) = \sigma^2 \sum_{j=1}^{p} (d + \lambda_j)^{-1}$. ¿Cuándo sería conveniente incorporar esta información adicional para reducir $\sum_{j=1}^{p} Var(\widehat{\beta}_j)$?

9. Sean $R^2_{a,k}$ y $R^2_{a,p}$ los R^2 ajustados de los modelos (1) y (2) respectivamente

$$E(Y) = 1 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} \qquad (1)$$

$$E(Y) = 1 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1} \qquad (2)$$

donde $p < k$.

a) Pruebe que $1 + \dfrac{C_p - p}{n - p} = \dfrac{1 - R^2_{a,p}}{1 - R^2_{a,k}}$, con lo cual $R^2_{a,p} > R^2_{a,p'}$ no es equivalente a $C_p < C_{p'}$.

b) Pruebe que $C_k = k$.

c) Sea F el estadístico del test F para testear la hipótesis de de que el modelo (2) es válido frente a la alternativa de que el modelo (1) es cierto. Pruebe que

$$F = \frac{1}{n - p}(C_p - 2p + k).$$

10. La Tabla 4 (archivo llamadas.txt) muestra los datos correspondientes al número total de llamadas telefónicas internacionales (en decenas de millón) entre 1950 y 1973 registradas en el Belgian Statistical Survey publicada por el Ministerio de Economía de Bélgica.

a) Ajuste por mínimos cuadrados un modelo lineal simple usando como variable explicativa a *year* y como respuesta a *calls*. Calcule los residuos estandarizados y

studentizados, el leverage, la distancia de Cook y los Dfits para el modelo ajustado. Calcule los puntos de corte para cada medida y realice los gráficos que le parezcan apropiados con estas medidas para identificar posibles outliers y/o puntos influyentes.

b) Realice un scatterplot de *year* vs. *calls* en el que se grafique además la recta obtenida por mínimos cuadrados. ¿Qué le parece el ajuste obtenido? ¿Le parece que lo que se observa en el gráfico es bien reflejado por las medidas calculadas en a)? Si la respuesta es negativa, ¿cuál puede ser el motivo?

c) Para el mismo modelo, ajuste los datos mediante un M–estimador basado en la función ψ de Huber con constante $c = 1.345$. Agregue al gráfico de b) la recta obtenida. ¿Nota diferencias?

d) Idem c) utilizando la función bicuadrada con constante $c = 4.685$.

e) ¿Hay diferencias entre los dos ajustes realizados por los M–estimadores? Si las hay, ¿a qué se pueden deber?

Imprima el vector de pesos de cada estimador y observe cuál es el peso que se le da a cada observación.

11. Los datos de la Tabla 5 (stars.txt) forman el diagrama de Hertzsprung–Russell del cluster de estrellas CYG OB1, que contiene 47 estrellas en la dirección de Cygnus. *Ltemp* es el logaritmo de la temperatura en la superficie de la estrella y L es el logaritmo de intensidad lumínica.

a) Ajuste por mínimos cuadrados un modelo lineal simple usando como variable explicativa a *Ltemp* y como respuesta a L. Calcule los residuos estandarizados y studentizados, el leverage, la distancia de Cook y los Dfits para el modelo ajustado. Calcule los puntos de corte para cada medida y realice los gráficos que le parezcan apropiados con estas medidas para identificar posibles outliers y/o puntos influyentes.

b) Realice un scatterplot de *Ltemp* vs. L en el que se grafique además la recta obtenida por mínimos cuadrados. ¿Qué conclusiones obtiene? ¿Le parece compatible este resultado con los del item a) ?

c) Para el mismo modelo, ajuste los datos mediante los M–estimadores basados en la función ψ de Huber con constante $c = 1.345$ y la función bicuadrada con constante $c = 4.685$.

Agregue al gráfico de b) las rectas obtenidas. ¿Nota diferencias? ¿Qué puede estar pasando?

Analice el vector de pesos de cada M–estimador y observe cuál es el peso que se le da a cada observación.

d) Calcule el LMS–estimador (estimador de mínima mediana de cuadrados). Superponga al scatterplot resultante de los items anteriores la recta ajustada por este método.

¿Qué observa?

e) Realice boxplots paralelos de los residuos estandarizados correspondientes a todos los estimadores calculados. Para los estimadores robustos una forma de estandarizar los residuos puede ser dividir al residuo crudo por (1.4826 Mad(residuos)), que es una escala robusta de los residuos. Identifique los que son marcados como outliers.

f) Recalcule el estimador de mínimos cuadrados eliminando aquellos puntos cuyos residuos son marcados como outliers en el ajuste realizado por el LMS–estimador. Superponga la recta obtenida en el scatterplot. ¿Qué observa?

12. En la Tabla 6 (stack.txt) se presentan los datos conocidos como *Stackloss Data* ampliamente tratados en la literatura. Los 21 datos corresponden a una muestra real y describen la operación de una planta de oxidación de amoníaco a ácido nítrico. La variable de respuesta Y es el stackloss, $X1$ es la tasa de operación, $X2$ es la temperatura del agua de enfriado y $X3$ es la concentración de ácido.

Resumiendo el análisis realizado por diversos autores, puede decirse que las observaciones 1, 3, 4 y 21 fueron clasificadas como outliers, mientras que la observación 2 también fue calificada como outlier moderado por algunos de estos autores.

a) Realice un ajuste por mínimos cuadrados, realice un gráfico de residuos estandarizados vs. número de observación y verifique si el valor absoluto de algún residuo estandarizado es mayor que la cota 2.5.

b) Repita a) utilizando los estimadores robustos usados en el ejercicio anterior y la forma sugerida de estandarizar los residuos obtenidos. ¿Alguno de estos gráficos le permite coincidir con la clasificación de ouliers presentada en la bibliografía ?

13. Usando la Tabla 7 identifique las principales causas de la colinealidad.

14. Utilizando los datos del ejemplo de Biomasa (Tabla 6 de la Práctica 3 y archivo biomasa.txt),

a) Calcule para todos 31 subconjuntos basados en las 5 variables independientes: K, $SODIO$, PH, SAL y ZN, el C_p de Mallows, el R^2 y el R^2 ajustado. Elija los subconjuntos que serían los mejores candidatos según cada criterio. Compare las elecciones realizadas por los distintos métodos. Realice un gráfico de p vs. C_p y observe cuáles son los subconjuntos con C_p pequeño y cercanos a p.

b) Use el procedimiento Stepwise con la opción "forward" para elegir el mejor modelo. Describa el test F que se realiza en cada paso (es decir cuál es la hipótesis nula y cuál es el modelo en cada paso). Verifique su respuesta realizando el test adecuado.

c) Use el procedimiento Stepwise con la opción "backward" para elegir el mejor modelo. Describa el test F que se realiza en cada paso (es decir cuál es la hipótesis nula y cuál es el modelo en cada paso).

d) Use el procedimiento Stepwise con la opción "Efroymson" para elegir el mejor modelo. Describa el test F que se realiza en cada paso (es decir cuál es la hipótesis nula y cuál es el modelo en cada paso).

Compare la elección elegida automáticamente con las resultantes del ítem a).

15. Para los datos del ejercicio 18 de la Práctica 2 (peak.txt) transformados con logaritmo,

a) calcule, con la función leaps, para los mejores subconjuntos que selecciona por default basados en las 9 variables independientes, el C_p de Mallows, el R^2 y el R^2 ajustado. ¿Cuáles son los mejores candidatos de acuerdo a cada criterio ? ¿Son muy diferentes sus elecciones?

b) Use el procedimiento Stepwise con la opción Efroymson para elegir el mejor modelo automáticamente. Compare este modelo con los obtenidos en el ítem a). ¿Cómo se pueden interpretar las diferencias observadas?

16. Los datos de la Tabla 8 (webster.txt) fueron generados por Webster, Gunst y Mason (1974). Los generaron de manera tal que $\sum_{j=1}^{4} x_{ij} = 10$ para las observaciones 2 a 12 y $\sum_{j=1}^{4} x_{ij} = 11$ para la observación 1. Las variables x_5 y x_6 fueron generadas con distribución normal, mientras que la respuesta Y satisface el modelo:

$$Y = 10 + 2x_1 + x_2 + 0.2x_3 - 2x_4 + 3x_5 + 10x_6 + \epsilon$$

donde $\epsilon \sim N(0, 1)$.

a) Calcule la matriz de correlación de las variables independientes $x_i, i = 1, 6$ ¿Le parece que están altamente correlacionadas?

b) Calcule el estimador de mínimos cuadrados y observe cuáles son los estimadores de los parámetros que son significativos con nivel 0.05.

c) Calcule los Factores de Inflación de la Varianza (VIF) para este ejemplo. ¿Qué le sugieren?

d) Calcule los índices de condición para X_s, la matriz de diseño escalada (es decir, luego de dividir a cada columna de la matriz de diseño por su norma). ¿Qué le sugieren ?

e) A partir de la matriz X_s realice la descomposición proporcional de la varianza de los estimadores de los parámetros y analícela. ¿Cuáles son sus conclusiones?

Tabla 4. Llamadas

obs	*year*	*calls*
1	50	4.4
2	51	4.7
3	52	4.7
4	53	5.9
5	54	6.6
6	55	7.3
7	56	8.1
8	57	8.8
9	58	10.6
10	59	12.0
11	60	13.5
12	61	14.9
13	62	16.1
14	63	21.2
15	64	119.0
16	65	124.0
17	66	142.0
18	67	159.0
19	68	182.0
20	69	212.0
21	70	43.0
22	71	24.0
23	72	27.0
24	73	29.0

Tabla 5. Stars

obs	*Ltemp*	*L*	obs	*Ltemp*	*L*
1	4.37	5.23	24	4.49	4.85
2	4.56	5.74	25	4.38	5.02
3	4.26	4.93	26	4.42	4.66
4	4.56	5.74	27	4.29	4.66
5	4.30	5.19	28	4.38	4.90
6	4.46	5.46	29	4.22	4.39
7	3.84	4.65	30	3.48	6.05
8	4.57	5.27	31	4.38	4.42
9	4.26	5.57	32	4.56	5.10
10	4.37	5.12	33	4.45	5.22
11	3.49	5.73	34	3.49	6.29
12	4.43	5.45	35	4.23	4.34
13	4.48	5.42	36	4.62	5.62
14	4.01	4.05	37	4.53	5.10
15	4.29	4.26	38	4.45	5.22
16	4.42	4.58	39	4.53	5.18
17	4.23	3.94	40	4.43	5.57
18	4.42	4.18	41	4.38	4.62
19	4.23	4.18	42	4.45	5.06
20	3.49	5.89	43	4.50	5.34
21	4.29	4.38	44	4.45	5.34
22	4.29	4.22	45	4.55	5.54
23	4.42	4.42	46	4.45	4.98
			47	4.42	4.50

Tabla 6. Datos de Stackloss

obs	$X1$	$X2$	$X3$	Y
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

Tabla 7

autovalores	número de condición	$var(b_0)$	$var(b_1)$	proporciones de $var(b_2)$	$var(b_3)$	$var(b_4)$	$var(b_5)$
5.7698	1.0	0.0009	0.0033	0.0000	0.0001	0.0001	0.0000
0.1313	43.944	0.0007	0.7599	0.0008	0.0004	0.0003	0.0001
0.0663	87.086	0.2526	0.0484	0.0038	0.0027	0.0057	0.0001
0.0268	215.796	0.3761	0.0003	0.0012	0.0019	0.0695	0.0000
0.00573	1006.793	0.0402	0.1705	0.0634	0.2585	0.0004	0.0001
0.00017	34596.0	0.3295	0.0176	0.9307	0.7364	0.9241	0.9997

Tabla 8. Datos de Webster et al.

obs	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$	Y
1	8	1	1	1	0.541	−0.099	10.006
2	8	1	1	0	0.13	0.07	9.737
3	8	1	1	0	2.116	0.115	15.087
4	0	0	9	1	−2.397	0.252	8.422
5	0	0	9	1	−0.046	0.017	8.625
6	0	0	9	1	0.365	1.504	16.289
7	2	7	0	1	1.996	−0.865	5.958
8	2	7	0	1	0.228	−0.055	9.313
9	2	7	0	1	1.38	0.502	12.96
10	0	0	0	10	−0.798	−0.399	5.541
11	0	0	0	10	0.257	0.101	8.756
12	0	0	0	10	0.44	0.432	10.937